# CHAPTER 13
# RELIABILITY

## STUDENT-LEVEL RELIABILITY

Although accountability decisions apply at the school level, student-level results are reported to parents. It is useful, therefore, to examine reliabilities at that level. Table 13-1 presents student-level coefficient alpha for open-response items for Accountability Cycle 3 by grade, subject, and year. Coefficient alpha was computed by form for common and matrix items combined. Median and range alpha values were computed across all 12 forms. These values are based on data contributed by students who were eligible to complete open-response tests and who were present on the day of testing. Absence was defined as an observation in which all common and matrix items were blank. When an observation included at least one response, zeros corresponding to any blank items were entered in the computation of coefficient alpha. The responses of absent students (all blanks) were excluded to avoid overestimation of score reliability.

**TABLE 13-1**
**OPEN-RESPONSE TEST RELIABILITIES**
**COEFFICIENT ALPHA BY SUBJECT AND GRADE**

| Grade | Subject | 1995 | | 1996 | | 1997 | | 1998 | |
|---|---|---|---|---|---|---|---|---|---|
| | | Median[1] | Range | Median[1] | Range | Median[1] | Range | Median[1] | Range |
| 4/5 | Reading | .81 | .80-.85 | .81 | .79-.83 | .80 | .77-.82 | .81 | .80-.84 |
| | Mathematics[2] | .82 | .81-.84 | .81 | .80-.83 | .76 | .74-.77 | .82 | .81-.83 |
| | Science | .76 | .72-.78 | .75 | .73-.77 | .72 | .71-.76 | .71 | .68-.73 |
| | Social Studies[2] | .80 | .78-.82 | .83 | .80-.84 | .77 | .73-.79 | .76 | .73-.79 |
| | Composite | .93 | .93-.94 | .93 | .93-.94 | | | | |
| 7/8 | Reading[3] | .85 | .82-.86 | .85 | .84-.86 | .85 | .82-.86 | .85 | .84-.86 |
| | Math | .84 | .81-.86 | .85 | .82-.86 | .79 | .78-.81 | .79 | .78-.80 |
| | Science[3] | .81 | .77-.82 | .80 | .78-.82 | .77 | .73-.79 | .79 | .76-.82 |
| | Social Studies | .88 | .86-.89 | .86 | .84-.88 | .85 | .82-.86 | .85 | .83-.86 |
| | Composite | .94 | .94-.95 | .95 | .94-.95 | | | | |
| 11 | Reading | .85 | .84-.87 | .85 | .83-.86 | .86 | .83-.87 | .87 | .84-.88 |
| | Mathematics | .87 | .85-.88 | .84 | .83-.85 | .82 | .79-.83 | .86 | .85-.87 |
| | Science | .84 | .80-.87 | .82 | .80-.84 | .80 | .76-.82 | .81 | .78-.83 |
| | Social Studies | .87 | .87-.89 | .85 | .83-.87 | .86 | .83-.88 | .89 | .86-.90 |
| | Composite | .96 | .95-.96 | .95 | .94-.95 | N/A[4] | N/A | N/A | N/A |

[1] Median coefficient alpha based upon common and matrix items across 12 forms.
[2] Grade 4 in 1995 and 1996; Grade 5 in 1997 and 1998.
[3] Grade 8 in 1995 and 1996; Grade 7 in 1997 and 1998.
[4] N/A, Not Available.

The Kentucky Department of Education (KDE) advises against making student-level decisions based on individual test scores alone. However, KIRIS open-response test reliabilities compared favorably with reliabilities from other tests used to make student-level decisions. KIRIS composite score reliability for 1995, 1996, 1997, and 1998 was

comparable to reliabilities published for the ACT Composite, and individual KIRIS subject area reliabilities were similar to ACT and CTBS subject area reliabilities.

It should be noted that using coefficient alpha probably underestimated score reliability insofar as item raw scores were the basis for the computation. A seven or eight item test having a single, relatively difficult item is likely to yield a lower coefficient alpha than a test of comparable length featuring items of essentially uniform and moderate difficulty. The fundamental scaling method used by KIRIS employed a logistic model. The use of item response theory took into account differences in item difficulty not reflected in the computational use of raw scores.

Coefficient alpha did not take into account variability in student-level scores arising from the use of multiple scorers. As indicated in Chapter 8, however, the effect of having different scorers appeared minimal even at the student level: correlations among ratings assigned by two scorers to the same student's responses in a given content area ranged from .87 to .99. This, and the fact that scoring was monitored through a quality assurance process, suggested that the effect of scorers on a given student's scores was very small, probably negligible, for the KIRIS composite.

At the same time, coefficient alpha could be inflated by a within-student scorer effect if one person scored all 7 of a student's responses in a given content area. To test this possibility, coefficient alpha was recomputed using the sample of grade 11, 1996, students whose papers were scored twice. See Chapter 8, page 8-8 for the description of the procedures that led to double scoring. Median coefficient alpha, across forms, for scores of odd-numbered items assigned by the first scorer and scores to the even numbered items assigned by the second scorer, were comparable to results presented in Table 13-1; two content areas yielded slightly lower alpha, two others yielded slightly higher alpha, while the composite score alpha was unchanged. Thus, scorer effects did not appear to inflate coefficient alpha estimates appreciably.

Because Arts & Humanities and Practical Living/Vocational Studies items were used only for purposes of school-level score reporting, no estimates of student-level score reliability were provided for these subject areas. Coefficient alpha was not computed for Writing and Mathematics Portfolio scores, since each portfolio receives only one score. At least two scores per student are necessary to compute a student-level reliability statistic. See Chapter 12 for portfolio scoring consistency data.

## SCHOOL-LEVEL RELIABILITY

The main concern for school-level reliability is *consistency* of results. Consistency of results with respect to schools is more complex involving multiple sources of inconsistency. First, to what extent do the results obtained in a given year on the basis of the particular test items used that year generalize to a theoretical pool of other items that might have been selected. Secondly, to what extent do the particular students within each school in an accountability grade generalize to a theoretical pool of other students who might have attended an accountability grade?

School-level reliability is concerned with the following question: if an alternative set of test items from the same domain was administered to an alternative set of students attending the school, to what extent would the school have obtained the same accountability rating for that particular accountability cycle? School-level results raise the important question of *decision consistency,* critical, because of the consequences of school-level decisions.

The KIRIS accountability system required an accountability index for each accountable school and district in the state. The accountability index was computed as the weighted average of eight sub-indices measuring specific constructs. Seven of the sub-indices reflected performance in cognitive areas (Reading, Writing, Mathematics, Science, Social Studies, Arts & Humanities and Practical Living/Vocational Studies). These sub-indices were based on results from open-response testing and/or portfolios. The eighth index was a combination of the noncognitive measures used to assess the effectiveness of schools. These include attendance, dropout rate, retention, and successful transition to adult life.

Because schools were judged by their improvement over time, rather than on the basis of their status at any one time, a gain score was used as the basis for reward and assistance decisions. The gain score was based on biennial school accountability indices calculated over a four-year accountability cycle. The computation involved first, setting a baseline index, i.e., the average of results generated during the first and second years of the cycle; and second, comparing this *baseline* average to the *growth* average, the index produced by averaging the two succeeding years of the cycle. The difference between the baseline and growth averages comprised the *gain*. It was evaluated with respect to the improvement goal established after the first two years of the cycle. See chapter 10, page 1 for the formula used to compute the improvement goal.

A reliability index is needed to interpret the accountability index gain score. A common approach to the computation of reliability requires the testing and retesting of students. This was neither reasonable nor feasible in the context of KIRIS. It was therefore decided to estimate accountability gain-score reliability by using a simulation study.

The strategy of the study described below was to estimate variability in scores due to measurement error and sampling error in the first stage, using schools sufficiently large to support a statistical technique known as sampling with replacement. This first stage yielded an estimate of gain score standard error. The second stage used actual school score and size data to simulate variability in scores for each year of the accountability cycle one hundred times, comparing the school accountability classification obtained in each of the hundred replications with the classification assigned in reality. The second stage yielded an estimate of decision consistency.

## ACCOUNTABILITY INDEX GAIN SCORE STANDARD ERROR OF MEASUREMENT

To estimate the standard error of measurement for the KIRIS accountability index gain score, a computer simulation was performed using actual school results for Accountability Cycle 2. This study was not replicated during Accountability Cycle 3. To simulate the variability in observed scores associated with the students enrolled in an accountability grade in a given school in a particular year (compared to those who might have enrolled if, for example, they had been a little younger or a little older), multiple simulation sets of students were created. This was accomplished by drawing (with replacement) samples from each school. This provided a simulation of sampling error. To simulate the variability in obtained scores associated with the items that appear on KIRIS tests each year relative to the items that could appear, each time a student was selected to be in the simulation study, the student's observed score was re-sampled, or changed, to simulate the fact that the student would probably score somewhat differently on a different set of test items. (For portfolio scores, this re-sampling of student score was accomplished by sampling from the possible scores that might have been obtained if a different teacher had scored the portfolio.) This first kind of variability will be referred to as sampling error (or student sampling error), whereas the second kind will be referred to as measurement error. Note that variability in noncognitive indicators was simulated as well.

Using schools sufficiently large to provide for meaningful re-sampling of students, accountability index scores were computed for each school 50 times. To estimate sampling error and measurement error, pooled within-school variance estimates were computed across the 50 scores after the first step of the simulation (drawing multiple sets of students) and the second step (re-sampling or changing individual test scores for sampled students). As a check on the process, the second step was also conducted without the first step, i.e., individual test scores were changed 50 times for the same set of sampled students. This procedure yielded an estimate of the gain score standard error of measurement expected for each school size included in the study, i.e., 24, 36, 48, and 96. (Multiples of 12 were selected to accommodate student re-sampling within form, which was necessary to reflect the matrix-sampled design used in KIRIS.)

Gain score standard error of measurement figures for schools of size 96, over all four years of Accountability Cycle 2 are presented in Table 13-2. Larger gain score standard errors were found for smaller sample sizes, and were almost exactly proportional to the theoretically expected value, based on the inverse of the square of sample size. The sizes of the standard error of measurement due to sampling error were substantially greater than standard error due to measurement error (almost twice as large). This was consistent with previous results.

**TABLE 13-2**
**ESTIMATED ACCOUNTABILITY INDEX GAIN SCORE STANDARD ERROR OF MEASUREMENT**
**FOR SCHOOLS OF N = 96 STUDENTS**

| School Level | Sources of Score Variability Considered | | |
|---|---|---|---|
| | Sampling Error | Measurement Error | Measurement and Sampling Error[1] |
| Elementary | 1.49 | .76 | 1.42 (1.67) |
| Middle | 1.61 | .71 | 1.46 (1.76) |
| High | 1.52 | .71 | 1.43 (1.68) |

[1]The first figure given in this column provides the actual results obtained in the study in a two-step simulation of sampling error followed by measurement error, whereas values in parentheses show the result obtained through additive combination of separately obtained sampling and measurement error variance estimates.

As suggested by the footnote to Table 13-2, an unusual result was obtained in the study when pooled within-school variance results were computed from school scores obtained after applying both procedures, i.e., sampling of students (sampling error) and changing student scores (measurement error). A lower standard error of measurement was obtained across school scores when both measurement error and sampling error were incorporated, than when sampling error alone was considered.

This result is somewhat puzzling, although it may be a legitimate outcome of the performance level based scheme approach to the particular data in Accountability Cycle 2. That is, replications of schools yielding an unusual draw of students, such as a very strong or very weak group of students relative to the population, may yield scores that are drawn inward through the simulation of measurement error. Student scores assigned to the Novice performance level can only obtain a performance level higher than or equal to Novice in the simulation, and those assigned Distinguished can only obtain a performance level lower than or equal to Distinguished. Those assigned to Distinguished can vary in either direction in the simulation. In any case, a more conservative result (one suggesting a greater amount of error) was obtained through additive combination of variances. These standard error figures appear in parentheses in Table 13-2 as explained in the table footnote.

The procedures followed in the simulation study were described in greater detail in *Effects of Students and Tasks on Gain Scores Used in Complex School Accountability Decisions* (Dings & Kingston, 1996).

## ACCOUNTABILITY DECISION CONSISTENCY.

The preceding section discussed the computation of standard error of measurement estimates for accountability index gain scores for schools of the sizes simulated in the

study – 24, 36, 48, and 96. School sizes were held constant for the course of the four-year accountability cycle. Assessing decision consistency for schools varying in the number of students at the accountability grade over the four years required the consideration of three information elements: the obtained accountability index score for each year, the accountability grade enrollment, and an estimate of error for the accountability index. The first two information elements were a matter of record, whereas the third was obtained through application of results presented above.

Applying the above results was relatively straightforward. The application assumed that the standard error of measurement for a single year accountability index score was equal to the standard error of measurement for a gain score. The gain score was computed by subtracting the average of two single year indices from the average of two other single year indices. All other things being equal, doubling the number of years from one to two reduced standard error of measurement by a factor of the square root of 2; however, relying on a gain score increased standard error of measurement by the same factor. Therefore. it was reasonable to use figures presented in Table 13-2 as standard error of measurement estimates for a single year accountability index when 96 students were present in the accountability grade. Error estimates for other accountability grade enrollments were obtained through multiplication by the square root of 96 and division by the square root of the actual enrollment.

In estimating decision consistency, accountability index scores were computed for each of the four years of Accountability Cycle 2 as draws from a normal distribution. The mean of the distribution was assumed equal to the school accountability index mean for a given year, and the standard deviation equal to the standard error of measurement associated with accountability grade enrollment for that year. Results from each year were averaged appropriately (weighted by accountability grade enrollment) to obtain a baseline, a growth index, and a corresponding accountability decision for each of one hundred replications. Decision consistency was assessed by comparing results of the one hundred replications with the original accountability classification, which was assigned to each school.

Given that some uncertainty accompanies any measurement, perfectly consistent accountability decisions were not possible. However, the magnitude of the consequences of a decision should influence the level of consistency we require of the measurement used to make decisions. KIRIS results were used to assign cash rewards to schools whose student test scores exceeded improvement goals as well as to sanction schools whose scores declined substantially. To ensure the most efficient use of taxpayer dollars, we must be highly confident that a school's achievement has truly improved if it was to be rewarded. Similarly, to respect local control of schools and to avoid unwarranted sanctions, one must be highly confident that a school's achievement had truly declined, if it is to be designated a "school-in-crisis." (The rationale here is parallel to that applied to the College Board's Advanced Placement (AP) Examinations.)

Based upon concern for the expenditure of taxpayer dollars and the preservation of opportunity for local control of schools, the following two questions, and accompanying results, were important relative to Kentucky's accountability program:

1. What percentage of schools categorized as "Reward" truly improved their achievement?
   Based on re-sampling analyses for accountability cycle two, over 99 percent of the time, schools in the simulation originally assigned to the "Reward" classification obtained scores in the simulation in which the growth index had increased relative to accountability baseline**.** About 94 percent obtained scores in which the growth index had met or exceeded the improvement goal derived from the accountability baseline.

2. What percentage of schools categorized as "In Crisis" truly did not improve their achievement?
   Over 99 percent of the time, schools in the simulation originally assigned to "In Crisis" obtained scores in the simulation in which the growth index did NOT increase relative to, nor did it equal the accountability baseline.

Table 13-3 presents conditional probability percentages. The row headings, e.g. "In Crisis," group schools according to the 1996 accountability decisions. Column headings repeat these categories, reflecting the categories into which schools were assigned when reclassified by the simulation study. Consider, for example, the first cell of the first row of Table 13-3. This cell indicates that 0.866 of schools whose accountability classification was "In Crisis" were reclassified as "In Crisis" by the simulation study. The second cell in that row indicates that 0.134 of replications of schools originally classified as being "In Crisis" yielded an accountability classification of "In Decline" in the simulation study. These and the other results presented in Table 13-3 represented a tolerable level of decision inconsistency. The "Success" classification was the only inconsistently replicated accountability decision, a result which was not surprising when one considers that scores leading to a classification of "Success" encompass a very short range of the accountability gain score scale. That range included a single point, from just above "Improving" to just below "Rewards". The rationale for the "Success" classification was to provide a buffer between "Improving" which involved a very modest sanction (writing a school improvement plan) and "Reward." The remaining classifications yielded consistent decisions, with the two bounded categories ("Improving" and "Decline") showing results identical to their original decisions 0.81 of the time, and the unbounded categories ("In Crisis" and "Rewards") yielding such results 0.87 of the time.

It should be noted that decision consistency results were based on decisions for elementary, middle and high schools taken separately. They did not include decisions based on combined elementary/middle or middle/high school grades. Also, in keeping with KDE's position that school scores based on extremely small numbers of students are too vulnerable to sampling error to be used with the desired level of confidence for purposes of school accountability, classification consistency was based only on schools

of size twenty or larger in each of the four years of the Accountability Cycle 2. The school size of twenty was chosen somewhat arbitrarily for purposes of implementing this study. This decision should not be confused with or taken as an indication of KDE policy. Generally, there was less uncertainty about decisions involving large schools, and those showing exceptionally large score gains or declines. Standard error of measurement information may be used to quantify the level of uncertainty of a particular accountability decision relative to possible other possible decisions.

**TABLE 13-3**
**CONDITIONAL PROBABILITIES ILLUSTRATING ACCOUNTABILITY DECISION CONSISTENCY**

| Original Decision | Proportion of Original Decisions In Which Each Accountability Decision Occurred in Simulation | | | | |
|---|---|---|---|---|---|
| | In Crisis | Decline | Improving | Success | Reward |
| In Crisis | .866 | .134 | .000 | .000 | .000 |
| Decline | .038 | .813 | .148 | .000 | .000 |
| Improving | .000 | .058 | .813 | .069 | .061 |
| Success | .000 | .000 | .412 | .239 | .349 |
| Reward | .000 | .000 | .065 | .069 | .867 |

Note: Due to rounding, row percentages do not sum to 1.000 in every case. Column percentages are not intended to sum to 1.000, as the table provides conditional probabilities within each row.